UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Field, S. M., et al. (2019). When and Why to Replicate: As Easy as 1, 2, 3? *Collabra: Psychology,* 5(1): 46. DOI: https://doi.org/10.1525/collabra.218

ORIGINAL RESEARCH REPORT

When and Why to Replicate: As Easy as 1, 2, 3?

Sarahanne M. Field, Rink Hoekstra, Laura Bringmann and Don van Ravenzwaaij

The crisis of confidence in psychology has prompted vigorous and persistent debate in the scientific community concerning the veracity of the findings of psychological experiments. This discussion has led to changes in psychology's approach to research, and several new initiatives have been developed, many with the aim of improving our findings. One key advancement is the marked increase in the number of replication studies conducted. We argue that while it is important to conduct replications as part of regular research protocol, it is neither efficient nor useful to replicate results at random. We recommend adopting a methodical approach toward the selection of replication targets to maximize the impact of the outcomes of those replications, and minimize waste of scarce resources. In the current study, we demonstrate how a Bayesian re–analysis of existing research findings followed by a simple qualitative assessment process can drive the selection of the best candidate article for replication.

Keywords: replication; transparency; psychological science; Bayesian reanalysis

In 2005, Ioannidis published a theoretical article (2005), in which he argued that more than half of published findings may be false. The landmark mass replication effort of the Open Science Collaboration (henceforth OSC; 2015) gave empirical support for Ioannidis's claims a decade after they were made, but reported an even bleaker narrative. Only 36% of replication studies were successful in yielding a result comparable to that reported in the original article (more recent mass replication attempts have revealed similarly low reproducibility levels: Klein et al., 2018; Camerer et al., 2018). To put this finding in context: had all of the original results been true, a minimum reproducibility rate of 89% would be expected, according to the OSC (2015). These figures reflect the gravity of what is now known as the crisis of confidence, or replicability crisis, in science. Though the discussion began in psychology, reports of unsatisfactory reproducibility rates have come from many different fields in the scientific community (Baker 2016; Begley & Ioannidis, 2015; Chang & Li, 2018).

The literature has suggested a number of potential causes for poor reproducibility of research findings. One of the most obvious candidate causes is the publish or perish culture in academia (Grimes, Bauch, & Ioannidis, 2018), which describes the pressure on researchers to publish much and often in order to maintain their university faculty positions, or to move up the hierarchical 'ladder'. Another possible cause is the alarmingly high prevalence of QRPs (questionable research practises) in which researchers engage. HARKing (hypothesizing after the results are known), p-hacking (where one massages the data to procure a significant p-value) and the 'file drawer' problem (where researchers do not attempt to publish their null results) are all examples of QRPs (Kerr, 1998; John, Loewenstein, & Prelec, 2012; Rosenthal, 1979). They lead to a literature that is unreliable, and apparently in many cases (and often as a result), impossible to replicate.

Irrespective of the causes of the crisis of confidence, its consequence is irrefutable: scientific communities are questioning the veracity of many of the key findings of psychology, and are hesitant to trust the conclusions upon which they are based. A recent online *Nature* news story suggested that most scientific results should not be trusted (Baker, 2015). Research psychologists are asking whether science is "broken" (Woolston 2015); others have referred to the "terrifying unraveling" of the field (Aschwanden, 2016). Proposed solutions to this crisis of confidence have revolved around reviewers demanding openness as a condition to provide reviews (Morey et al., 2016), guidelines for more openness and transparency (Nosek et al., 2015), preregistration and registered reports (Dablander, 2017), and funding schemes directly aimed at replication (such as those of the Netherlands Organisation for Scientific Research: https://www.nwo.nl/en/news-and-events/news/2017/social-sciences/repeating-important-research-thanks-to-replication-studies.html).

These initiatives, while a first step in the right direction, only go so far to remedy the problem because they are preemptive in nature; only prescribing best practice for the *future*. They cannot help untangle the messy current

Department of Psychometrics and Statistics, Rijksuniversiteit Groningen, NL

Corresponding author: Sarahanne M. Field (sarahanne.field@gmail.com)

literature body we continue to build upon. The most direct way to get more clarity about previously reported findings may well be through replication (but see Zwaan, Etz, Lucas, & Donnellan, 2018, who discuss some caveats). Therefore, psychological science needs a way to separate the wheat from the chaff; a way to determine which findings to trust and which to disregard. Replication of existing empirical research articles is a practical way to meet this dire need. The need for replications introduces a second generation of complications related to interest in conducting replication studies: a flood of new replications of existing research can be found in the literature, and more are being conducted.

In theory, this up-tick in the number of replications being conducted is a good development for the field (especially given that up until recently, replication studies only occupied about 1 percent of the literature body: see Makel, Plucker, & Hegarty, 2012), however in practice, so much interest in conducting replications leads to a logistical problem: there exists a vast body of literature that *could* be subject to replication. The question is: how does one select which studies to replicate from the ever-increasing pool of candidates? Which replications retread already 'well-trodden ground', and which move research forward (Chawla, 2016)? These questions have serious practical implications, given the scarcity of resources (such as participants and time) in many scientific research fields.

Several recommendations to point us in the right direction exist in the literature already. A great number of these happen to be conveniently grouped as commentaries to Zwaan and colleagues' recent impactful article: 'Making Replication Mainstream' (2018). For instance, Coles, Tiokhin, Scheel, Isager and Lakens (2018), and Hardwicke and colleagues (2018) urge potential replicators to use a formalized decision making process, and only conduct a replication when the results of a cost-benefit analysis suggest that the benefits of such a replication outweigh the associated costs. Additionally, they emphasize considering other factors such as the prior plausibility of the original article's reported effects. Kuehberger and Schulte-Mecklenbeck (2018) argue against selecting replications studies at random and discuss potential biases that can emerge in the process of selecting studies to replicate. Little and Smith single out problems with existing literature as reasons for replication failure (such as weak theory measurement), which can be reasons for targeting some original studies for replication, over others. Finally, Witte and Zenker (2018) recommend replication only those studies which provide theoretically important findings to the literature. Coming from the opposite angle, Schimmack (2018) provides reasoning as to why to not replicate certain studies, which, naturally, is also useful in refining ones selection criteria for replication targets.

One could say that, broadly, there are three different facets to selecting replication targets, associated with the different information contained in a published article: statistical, theoretical and methodological. In the next two subsections, we first discuss statistical considerations and then theoretical and methodological considerations in turn.

Statistical Considerations

First, studies can be selected for replication when their claims require additional corroboration, based on the statistical evidence reported in the publication. This is a statistical approach to determining what should be replicated first. Null-hypothesis significance testing (or NHST) dominates the literature, meaning that the bulk of statistical testing involves reporting p-values.

Although there are numerous downsides to using NHST to quantify scientific evidence (for a discussion, see Wagenmakers, 2007), we focus on one key drawback here which relates directly to our discussion. The p-value only allows us to reject the null hypothesis: there is a single evidence threshold, meaning that we cannot use the p-value to gather evidence in favor of the null hypothesis, no matter how much evidence may exist for it. Given that it is unlikely that each study reporting an effect is based on a true main effect (Ioannidis, 2005), but that studies rarely use statistical techniques to quantify evidence for the absence of an effect, there is a mismatch in what we can conclude and what we want to conclude from our statistical inference (Haucke, Miosga, Hoekstra, & van Ravenzwaaij, 2019).

One alternative to quantifying statistical evidence with the conventional NHST framework is by means of *Bayes factors*. Throughout this paper, we will use a relatively diffuse *default* prior distribution for effect size to reflect the fact that we do not possess strong prior information (see also Etz & Vandekerckhove, 2016). In this paper we examine scenarios calling for a t-test. For such designs, one of the most prominent default specifications uses what is known as the Jeffreys-Zellner-Siow (JZS) class of priors. Development of these so-called JZS Bayes factors have been built on the pioneering work of Jeffreys (1961) and Zellner and Siow (1980). The JZS Bayes factor quantifies the likelihood of the data under the null hypothesis (with effect size $\delta = 0$) relative to the likelihood of the data under the alternative hypothesis. For a two-sided test, the range of alternative hypotheses is given by a prior on the effect size parameter δ, which follows a Cauchy distribution with a scale parameter $r = 1/\sqrt{2}$ (see, Rouder, Speckman, Sun, Morey, & Iverson, 2009, equation in note 4 on page 237). In terms of interpretation, a Bayes factor of $BF_{10} = 5$ means the data are 5 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. In comparison, a Bayes factor of $BF_{10} = 1/5$ (or the inverse of 5), means the observed data are five times more likely to have occurred under the null hypothesis than under the alternative hypothesis.[1]

The application of the JZS Bayes factor for a large-scale reanalysis of published results is not without precedent (Hoekstra, Monden, van Ravenzwaaij, & Wagenmakers, 2018). We build upon the work of Hoekstra and colleagues in taking the results of such a Bayesian reanalysis as a starting point for selecting replication targets (for a similar approach, see Pittelkow, Hoekstra, & van Ravenzwaaij, 2019).

So why not simply use p-values as our selection mechanism for existing statistical evidence? When NHST results are reanalyzed and transformed into Bayes factors,

the relationship between Bayes factors and p-values can be strong if the analyzed studies have mostly comparable sample sizes (Wetzels et al., 2011; Aczel, Palfi, & Szaszi, 2017). However, when studies have differing sample sizes, this relationship is no longer straightforward (for instance, see Hoekstra and colleagues (2018), who show that for non-significant findings the strength of pro-null evidence is better predicted by N than by the p-value, and that larger N studies are "more likely to provide compelling evidence", p. 6). Consider the following example for illustration.

We have two results of classical statistical inference:

Scenario 1: t(198) = 1.97, p = .05
Scenario 2: t(199998) = 1.96, p = .05

In both cases, the p-value is significant at the conventional alpha-level of .05, however due to the very different sample size in both scenarios, these two sets of results reflect very different levels of evidential strength. The Bayes factor, unlike the p-value, can differentiate between these two sets of results. Through the lens of the Bayes factor, scenario 1 presents ambiguous evidence: $BF_{10} = 0.94$ (i.e., the data is about equally likely to occur under the null hypothesis as under the alternative hypotheses). A Bayes factor for scenario 2 presents strong evidence in favor of the null: $BF_{10} = 0.03$ (i.e., the data is about 29 times more likely to occur under the null hypothesis than under the alternative hypothesis). Using the p-value as a criterion for which study to replicate would not differentiate between these two scenarios, whereas the Bayes factor allows us to decide that in case of Scenario 2, we have strong evidence that the null hypothesis is true (and so, arguably, no further replication is needed), whereas in case of Scenario 1, the evidence is ambiguous and replication is warranted.

In this paper, we apply a Bayesian reanalysis to several recent research findings, the end-goal being to demonstrate a technique one can use to reduce a large pool of potential replication targets to a manageable list. The Bayesian reanalysis is diagnostic in the sense that it can assist us in separating findings into three classes, or tiers of results: (1) results for which the statistical evidence pro-alternative is compelling (no replication is needed); (2) results for which the statistical evidence pro-null is compelling (no replication is needed); (3) results for which the statistical evidence is ambiguous (replication may be needed depending on theoretical and methodological considerations). We reiterate here that, crucially, p-values are unable to differentiate between results which belong in the second of these categorical classes, and those that belong in the third. The third class of studies will be carried into the next 'phase' of our demonstration, wherein we further scrutinize study results with ambiguous statistical evidence on theoretical and methodological considerations that might factor into the decision to replicate.

Theoretical and Methodological Considerations

Mackey (2012) provides some pointers on how one may select a replication target based on the theoretical content of a reported research finding. She suggests that in order to qualify as a 'candidate' for replication, a study should address theoretically important (for short, 'theoretical importance') and currently relevant research questions ('relevance'). A study also qualifies if it concerns studies in the field that are accepted as true in the field, but have yet to be sufficiently investigated ('insufficient investigation').[2] The theoretical approach will be explained as we describe it in a practical application later in the paper.

The last facet to selecting replication studies concerns methodological information. While many aspects of a study's methodology are highly specific to the paradigm of the article in question (e.g., the use of certain materials like visual stimuli), some elements of methodology can be discussed in general (e.g., sample size). As with the theoretical facet, methodology will be discussed in more detail during the later demonstration.

Outline

A replication study itself is beyond the scope of this paper, however we offer a demonstration of how the combined use of theory and Bayesian statistics can drive a methodical and qualitative approach to selecting replication targets in the psychological sciences. Additionally, we offer theoretical and methodological recommendations, in case such a replication were to be conducted. Please note that although the theoretical context and methodology of a study is important for selecting studies for replication, our demonstration focuses primarily on applying the Bayesian reanalysis to this challenge.

The remainder of this paper is organized as follows. In the method section, we share details of our treatment of the replication candidate pool in the reanalysis phase. We then describe the results of the initial selection process, before moving on to describing the qualitative phase of the filtering process. We make recommendations based on our selection process, for a fictional replication study. The article ends with our discussion, wherein we justify certain subjective choices we have made and consider philosophical issues, and share the limitations of our method.

Method

We extracted statistical details from articles in the 2015 and 2016 *Psychological Science* and performed a Bayesian reanalysis to make a first selection of which studies could be targets for replication, based on the evidential strength of the results reported. Once this initial selection was made, we further refined the selection based on the theoretical soundness of the conclusions drawn from the selected studies, and considered the support for the finding which exists in the literature already. The approach combined quantitative and qualitative methods: on the one hand, the initial selection was based on an empirical process, and on the other, the refinement of the selection was based on a process involving judgments of the findings in the context of the literature and theory. The process took the first author less than a working week to complete. Given that we provide the reader with the reanalysis code, and the spreadsheet with the necessary values to complete the reanalysis, we believe that attempts by others to use

our method for a similarly sized sample would not be any more time-intensive than our original execution.

Sample

All *Psychological Science* articles from 2015 and 2016 issues were searched for reported significant statistical tests (one–sample, paired, and independent t-tests), associated with primary research questions. As mentioned, we used statistical significance as our criterion for selecting results to reanalyze. All of the articles reporting t-tests to test their main hypotheses used p-values to quantify their findings. We extracted the t-values and other details required for the reanalysis (including N and p-value) for 30 articles which contained t-tests (the data spreadsheet which logs these details for each statistic extracted is on the project's Open Science Framework (OSF) page at https://doi.org/10.17605/OSF.IO/3RF8B).

Incomplete or unclear reporting practices posed a challenge in the first step of selecting which articles to reanalyze. Determining whether the executed tests were one- or two-sided was often difficult, as articles frequently failed to report the type of test conducted. Several articles which used t-tests as part of their main analysis strategy were ultimately not included in the reanalysis, as not all information was available (not even to the extent that we could reverse engineer other necessary details). One article, which reported two t-tests in support of their main finding, was excluded from the final reanalysis. Due to unclear reporting, we were unable to identify what the study's method entailed, and, therefore, how the reported results were reached. We explore the reporting problem in detail in the discussion section.

In total, from the 24 issues of 2015/2016 *Psychological Science*, 326 'research articles' and 'research reports' were manually scanned for studies in which a major hypothesis was tested using a t-test. Of these, 57 results were derived from 30 individual articles. Several articles reported more than one primary experimental finding which was analyzed using a t-test. Different approaches yielded judgments of whether or not a finding was of focal importance. First, if a specific finding was reported in the abstract, it would be selected (where possible). The rationale for this approach was that the abstract has only got space for documenting the most important results of the study, thus only key findings will be reported in it. A finding was also selected if somewhere in the article it was tested in a primary hypothesis, or was explicitly noted by the authors of the article as being important for the study's conclusions. Many articles reported several t-tests in support of a single broader hypothesis. In such cases we attempted to select the results which most directly supported the author's conclusions.

Descriptive Results

P-values, test statistics, sample sizes and test sidedness were collected for the purpose of the reanalysis. The p-values ranged in value; the largest was .047. The test statistics and sample sizes obtained also ranged greatly. The absolute test statistics ranged from 2.00 to 7.49. The range of the sample sizes is from $N = 16$ to $N = 484$. The distribution of study sample sizes is heavily right-skewed. The median for this sample is 54 – smaller than recent estimates of typical sample sizes in psychological research (Marszalek, Barber, Kohlhart, & Cooper, 2011).

In the Bayesian reanalysis, we converted reported information extracted from articles into Bayes factors, to assess the strength of evidence given by each result.[3] The Bayes factors range widely: 0.97 to 1.9×10^{10}, or approximately 19 billion. Almost half of them are between 1 and 5.

A clear negative relationship between the Bayes factors and the reported p-values is shown in **Figure 1**. Despite the nature of this relationship, some small p-values are associated with a range of Bayes factors (around the $p = .04$ mark, for instance). A positive relationship between Bayes factors and sample sizes can be seen in **Figure 2**. Unsurprisingly, larger sample sizes are generally associated with larger Bayes factors ($r = .71$), though it is not the case that large sample sizes are always associated with more compelling Bayes factors. For instance, many cases in the $N = 200$ region are associated with somewhat weak Bayes factors. In one case, the overall N of 30 converts to a Bayes factor of over 151,000, in another case, the overall N of 35 is associated with a Bayes factor of over 21,000.

Quantitative Target Selection

In this paper, we will make an initial selection based on those studies in tier 3: whose results yield only ambiguous evidence in relation to support for their reported hypotheses. For this purpose, we will judge such ambiguity, or low evidential strength, as when a study's BF_{10} lies between 1/3 and 3, which, by Jeffrey's (1961) classification system provides no more than 'anecdotal' evidence for one hypothesis over the other.

Using the BayesFactor package in R (Morey, Rouder, & Jamil, 2015), we calculated Bayes factors (BF) for each test statistic using the extracted test statistics, and other information gathered: p-values, test statistics, sample sizes and sidedness of the test. While the vast majority did not explicitly state that they were confirmatory, most results were presented as though they were. The code written for the analysis which is associated with the data spreadsheet can be found at the project's OSF page: https://doi.org/10.17605/OSF.IO/3RF8B.

The reanalysis revealed that the Bayesian reanalysis placed 20 results in evidence tier 3. One of these yielded a Bayes factor below 1 (0.97), which, by Jeffrey's classification system, demonstrates anecdotal pro-null evidence. The remainder of the results lie in tier 1. As we were only interested in those articles for which an effect was reported, no results falling in tier 2 (those with compelling pro-null evidence) exist in this dataset. The reanalysis has reduced the pool of results from 57 to 20 candidates for replication. We now move onto the next stage of target selection.

Qualitative Target Selection

Of the 20 results in tier 3, we select those demonstrating the weakest evidence for their effects. If there is an article for which many results fall in tier 3, these will also be

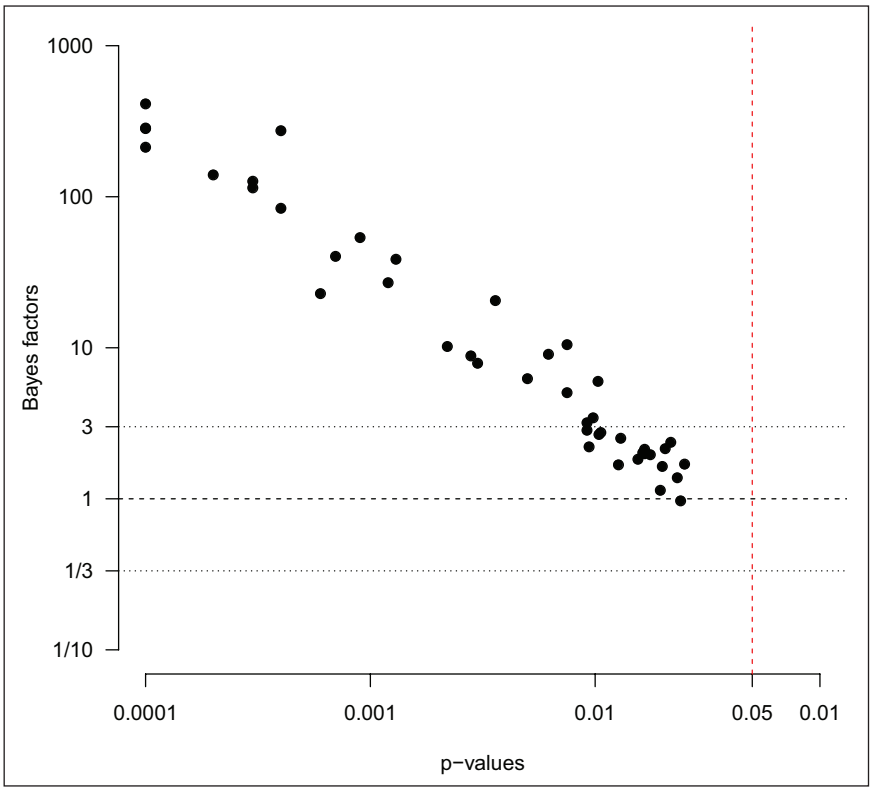

Figure 1: Scatterplot of Bayes factors and p-values plotted on a log-log scale. The horizontal dashed lines indicate Jeffreys' thresholds for anecdotal evidence (3, for pro-alternative cases, and the inverse for pro-null cases). The vertical red line demarcates the conventional significance level for p-values.

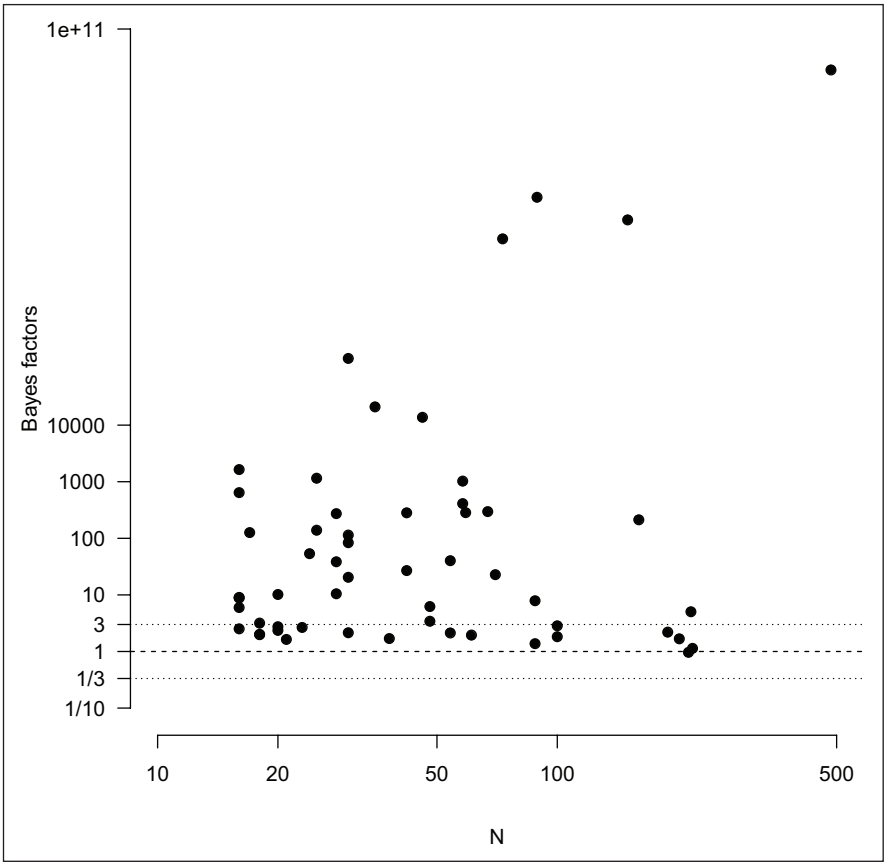

Figure 2: Scatterplot of Bayes factors and sample size plotted on a log-log scale. The horizontal dashed lines indicate Jeffreys' thresholds for anecdotal evidence (3, for pro-alternative cases; the inverse for pro-null cases). The cases in which we are interested for the reanalysis, those in tier 3, lie between the two finely dashed lines.

considered.[4] We will then conduct an assessment based on the qualitative criteria of Mackey (2012): theoretical importance, relevance, and insufficient investigation. Alongside Mackey's criteria, we consider the need for the finding in question to be replicated under different study conditions or with a different sample than the original (to establish the external validity of the effect in question), as well as replication feasibility (for instance, can this study be replicated by generally-equipped labs, or are more specific experimental set-ups necessary?). We will refer to the articles by the article number we have given them (the article and reanalysis details corresponding to these can be found in Appendix A; a full table of all the details can be found on the OSF page for this project, at https://doi.org/10.17605/OSF.IO/3RF8B).

The first to consider is the article revealed by the reanalysis to contain anecdotal pro-null evidence in one of its studies: article 8, from Dai, Milkman and Riis (2015). The authors of article 8 report on the so-called 'fresh start effect'. This effect refers to the use of temporal landmarks to initiate goal pursuit. More specifically, the authors' report supports claims that certain times of year (for instance, New Year's Eve) are especially potent motivators for starting new habits (such as working out, or eating more wisely). Although some evidence in this article is weakly pro-null (result 8a), one strike against naming article 8a as a suitable target for replication, is that the article contains a second result we reanalyzed (result 8b) which yielded a Bayes factor of 5.05 (constituting pro-alternative evidence: Gronau et al., 2017).[5]

In terms of Mackey's (2012) criteria, the study is difficult to judge as a replication target. Article 8's topic is theoretically important and certainly currently relevant: understanding the relationship between motivation and initiating healthy eating behavior is important for many reasons (for developing strategies to lowering the global burden of health due to preventable disease, for instance). However, the link between temporal landmarks and motivation has been demonstrated often and by different research groups (Peetz & Wilson, 2013; Mogilner, Hershfield, & Aaker, 2018; Urminsky, 2017), as well as in other studies by related groups (Dai, Milkman, & Riis, 2014; Lee & Dai, 2017), including a randomized clinical trial measuring adherence to medical treatment (Dai et al., 2017).

Although this phenomenon has been the subject of many different studies, and the content of article 8 lends itself to interesting replications in which one varies, for instance, the culture of the sample, existing literature in the area already demonstrates the effect in other cultures than the USA (e.g., Germany: Peetz & Wilson, 2013), it is not a clear replication target, in our assessment.

The majority of the remaining results in tier 3 show Bayes factors that are homogeneous in terms of their magnitude– for instance, half of the results have a Bayes factor between 1 and 2. Additionally, for articles with multiple reanalyzed studies, we see only one case in which each of these studies fell into tier 3. They may reflect one study of many in an article which overall, through other tests, provides strong evidence of a main effect. Both of these reasons render the majority of the sample less attractive as replication candidates.

Despite this, two articles (both featuring multiple low Bayes factors each) are potential targets.[6] We now commit these to the qualitative assessment to determine their suitability for replication, in no particular order.

One potential replication target is article 4 (Reinhart, McClenahan, and Woodman, 2015), in which the hypothesis that using mental imagery, or 'visualizing' can improve attention to targets in a visual search scene was tested. The authors recorded reaction times (RT) and event-related potentials (quantified as N2pc amplitudes, which reflect ongoing neural processes – in this case, attention) in response to the provided stimuli. They reported support for their hypothesis: imagining the visual search for certain targets did increase the speed at which participants focused on the specified targets (indexed by the ERP), before the motor response of pressing a button to confirm they had located the target. This article yielded three t-tests (each testing the experimental conditions on RT), which are of interest to us. We refer to them as results 4a through 4c, respectively. They appear in the results for the first experiment, which we judged to be a clear test of their primary hypothesis. Each of these t-tests correspond to a small Bayes factor. The RT tests correspond to Bayes factors of 3.19, 1.99 and 2.02, while the EEG tests yielded Bayes factors of 1.83 and 2.53. (the two other t-tests in the sets were not significant, thus are not of interest to us for the purposes of this reanalysis).

This article meets several of the qualitative criteria too. First, the topic is theoretically important and currently relevant. Training the brain for better performance has been gaining momentum in the past decade, partly prompted by several articles that support the positive link between video-gaming and improved mental performance in different cognitive domains (such as attention: Green & Bavelier, 2012, 2012). Exploring the link further with studies such as this can be beneficial to many areas of psychology and medicine (e.g., for working with patients of brain damage that are undergoing rehabilitation). Second, there is little supporting evidence for the link between visualization and improved attention; importantly, some of the literature aiming to reinforce the findings of article 4 contradicts it. For instance, the preregistered failed replication and extension of article 4's experiments conducted by Clarke, Barr and Hunt (2016) showed repeated searching – not visualization – improved attention. Other factors to consider are generalization and feasibility. The suitability of article 4 as a replication target is supported by fact that this article has already been a target for replication, and that that replication did not conclusively reinforce its conclusions. It is possible that this study should be weighted differently in the sample due to the previous replication. Indeed, one could numerically account for the evidence contributed by the existing replication (e.g., Gronau et al., 2017). We consider that to be outside of the scope of this paper.

Their sample for experiment one was comprised of adults between the ages of 18 and 35, with a gender split of 62% to 38% in favor of women. The findings of article 4

could benefit from a replication using a different sample: for instance, one with individuals from an older age range. Although age is not thought to impair neuroplasticity, older persons exhibit plasticity occurring in different regions of the brain than younger persons influencing the mechanisms underlying visual perceptual learning (Yotsumoto et al., 2014), which may influence their response to the stimuli presented in the experiments in the article. This has implications for the generalizability of the results. Another potentially important factor for consideration is gender. A recent review article by Dachtler and Fox (2017) reports clear gender differences in plasticity that are likely to influence several cognitive domains (including learning and memory), due to circulating hormones such as estrogen, which are known to influence synaptogenesis. To summarize, we find article 4 to be suitable as a replication candidate. Specifically, some of its findings could benefit from external reinforcement in the form of a conceptual replication in which factors such as age and gender are taken into consideration. Further, the results may benefit from a more in-depth exploration into the effect of searching versus visualization on attention.

Another replication target that our sample yielded is article 12: Kupor, Laurin and Levav (2015). Mentioned above, all reanalyzed results of this article (i.e., a through c) fall into tier 3. Article 12 (which includes 5 studies, each with sub-studies), explores the general hypothesis that reminders of God increase risk-taking behavior. In study 1, which this reanalysis focused solely on (as it most directly tested the key hypothesis), four sub-studies are identified: 1a, 1b, 1c and 1d. The first three contain t-tests, while the fourth contains a chi-square test. We consider only the results of 1a through c (12a through c) for the current reanalysis.

In the study corresponding to result 12a, participants performed a priming task involving scrambled sentences. Half the participants were primed with concepts of God, by way of exposure to words such as "divine" (p. 375). The other half, which forms the control group, were exposed only to neutral words. Once participants were primed, they completed a self-report risk-taking scale which was explained to participants as being an unrelated study. This scale revealed their likely risk-taking behavior in a one to five Likert scale. In the study yielding result 12b, following the manipulation, participants described the likelihood that they would attempt a risky recreational task that they had described themselves at an earlier point. In the study corresponding to result 12c, participants were tested on their interest in risk-taking via a behavioral measure, once they were primed in the first phase of the experiment. In each of these three experiments, participants primed with concepts of God reported or behaved as predicted: more predisposed to risk-taking than their neutrally-primed counterparts. Despite these three experiments yielding significant p-values, the reanalysis revealed three Bayes factors all suggesting the evidence is ambiguous: 1.96, 1.68 and 1.83, respectively for results 12a–c.

We now assess article 12 on the qualitative factors we described earlier. First, we consider the theoretical importance and current relevance of this article. Given that the majority of the world identifies as being religious (84%, according to recent statistics: Hackett, Stonawski, Potančoková, Grim, & Skirbekk, 2015), understanding the role of religion in moderating behavior is important, to say the least. According to the authors of article 12, behavior modification programs such as those employed for drug and alcohol rehabilitation use concepts of God and religion as a tool to reduce delinquent behavior. While this topic has attracted the attention of several research groups globally (meaning the article does not naturally meet the 'insufficient evidence' criterion), the reanalyzed results in article 12 go against the majority of this body of work: "... we propose that references to God can have the opposite effect, and increase the tendency to take certain types of risks" (p. 374), and do not seem to have direct strong support in the literature as yet (a paucity of indirect support can be found, e.g., Wu & Cutright, 2018).

In assessing the characteristics of article 12's sample, some details indicating the suitability of article 12 for replication come to light. First, article 12 reports using Amazon's Mechanical Turk online workforce, which is comprised of approximately 80% U.S.-based workers, and 20% Indian workers. Given that the majority of the Mechanical Turk workers are from the U.S., and the overwhelming majority of the U.S. reports being affiliated with Christianity, we expect that the majority of this sample respond with a mindset of trusting in a God which is thought to intervene on the behalf of the faithful, responding to prayers for things like healing, guidance and help with personal troubles. The results of article 12 might be very different if the participant pool contained mostly practitioners of Buddhism (for example), as Buddhism emphasizes the importance of enlightenment (when an individual achieves an understanding of life's truth), and personal effort, rather than the intervention of a divine being (which is relevant given that feelings of security are thought to increase willingness to engage in certain behaviors: p. 374).

The age of article 12's sample is also relevant to their results, considering that the majority of workers (>50%) were born in the 1980s. Recent polls indicate that younger individuals across Europe, the USA and Australia are less religious than their older counterparts (Harris, 2018; Wang, 2015; Schneiders, 2013), meaning that a successful replication of article 12's results with a predominantly aged population (as opposed to the mean ages of 23, 31 and 34 years, reported in the article) would demonstrate the generalizability of the finding that God-priming increases risk taking.[7] Another possibility also relates to age – perhaps the effect is greatly decreased in aged persons, simply by virtue of maturity: Risk-taking, even for rewards, decreases as a function of age (Rutledge et al., 2016).

Our reanalysis of article 12's results, in conjunction with other methodological and theoretical criteria considerations heavily underlines this replication candidate as a promising target, reporting results that are in need of independent corroboration. We recommend a direct, or pure replication, such that the findings exactly as they are presented can be verified. In addition, we recommend a

conceptual replication in which significant changes to the characteristics of the sample are made (e.g., as mentioned, on the basis of the participants' ages and religions).

Discussion

In this paper, we performed a large scale reanalysis of the results of a selection of articles published in *Psychological Science* in the years of 2015 and 2016 for which primary research findings were quantified by t-tests. Reanalyzing these results narrowed the pool of potential replication targets from 57 to 20 candidates. The Bayes factors for these candidate studies were between 0.97 and 2.85. To further our demonstration, we selected three articles, and subjected them to the second phase of the selection process, involving qualitative assessment. The qualitative process revealed that two of these articles are suitable for replication: their findings are theoretically important and relevant, but the literature largely lacks direct corroborating evidence for the claims thus far. It revealed that the results could benefit from changes to the magnitude of the samples, and that several variables should be included in conceptual replications to help generalize the reported results beyond the original articles.

A set of replications for articles 4 and 12 could first provide support for the existence of an effect, given the results of the Bayesian reanalysis. Once an underlying true effect is found to likely exist via a direct replication, further conceptual replications could be designed to explicitly explore other cohorts to better establish the generalizability of the findings beyond the original experimental cohort. In the case of article 4, specifically targeting participants of certain age groups may be beneficial to help determine the malleability of the effect across the lifespan. For article 12, targeting specific religious groups may assist in helping establish whether the God priming effect extends to other religions for which God is not a figure directly associated with intervention. These conceptual replications could also feature designs which vary from the originals – for instance, a replication of article 4 could feature a design in which gender is a blocking variable, or even included as a variable of interest.

Replications for both articles should contain much larger sample sizes, to help eliminate issues of reliability. In order to conduct a compelling replication study, one may need a sample size greater than that in the original study, depending on how large the sample is in the original study. Low experimental power produces some problems with reliability of original findings, leading to poor reproducibility even when other experimental and methodological conditions are ideal, which they rarely are (Button et al., 2013; Wagenmakers & Forstman, 2014).

A simulation by Button and colleagues (2013) demonstrates an argument against the common misconception that if a replication study has a similar effect size to the original, the replication will have sufficient power to detect an effect. They show that "... a study that tries to replicate a significant effect that only barely achieved nominal statistical significance (that is, $p \sim 0.05$) and that uses the same sample size as the original study, will only achieve ~50% power, even if the original study accurately estimated the true effect size" (p. 367). This indicates that in order to obtain sufficient power (say, $1-\beta = .8$) for a medium effect size in a replication study, the original sample would need to be more than doubled. In terms of the sample size in question, this indicates an increase from $N = 105$ to $N = 212$ for each of the replication studies.

Choice Justifications

Prior Choice

Though we do not want to rehash decades of debate about prior selection, our use of a Bayesian approach in our reanalysis stage, necessitates a brief discussion on our choice of prior. We have chosen to use the default prior – the Cauchy – in the BayesFactor package. This choice is suitable for our goals for a few reasons (and we recommend that the typical user use the package defaults for the same reasons). First, the Cauchy prior's properties make it an ideal choice for a weakly informative prior based on 'general desiderata' (Jeffreys, 1955). Second, even if we did want to use a subjective prior, the most obvious approach to doing so would yield unreliable results. Using the existing literature on an effect to inform one's prior choice would be a poor idea due to publication bias. Other factors exist that complicate subjective prior use. For instance, the existing literature on a particular phenomenon might be conflicting (in which case, the 'right' subjective prior might not exist), or may be very sparse (in which case little information would be available to adequately inform the prior). This being said, there are potential users of our method that may have sufficient expertise to navigate this complex situation and wish to select an alternative to the Cauchy prior. We refer such users to Verhagen and Wagenmakers (2014) or to Gronau, Ly and Wagenmakers (2019), both of which deal with Bayesian t-tests with explicit prior information available.

Selection Based on Significance

We used statistical significance as the criterion for selecting results for the Bayesian reanalysis. One may wonder why we have not chosen to inspect the claims of the non-existence of an effect based on a non-significant p-value. We have two reasons for using statistical significance (that is, when original article authors used statistical significance to justify their claims). First, although we believe statistical significance is hardly diagnostic of a true effect, the lack of statistical significance being related to no effect is even more complicated. If one were to try to replicate a non-significant result, what would the result say of the original effect? This problem does not exist for, say, an original study with a strong pro-null Bayes factor result, as the Bayes factor allows us to actually quantify pro-null evidence.

Finally, some applications of our method could be constrained by the capabilities or resources of replicating labs – not all suitable replication candidates can be replicated by all interested parties, as shown in our description above. The study of article 4 is worthwhile as a replication target and warrants further investigation, however it requires specialized equipment and specific

expertise to be recreated, and is therefore only feasible for select labs to seriously attempt. On the other hand, article 12 features a less specialized set of materials that could be recreated by a research group using easily-accessible university provided software (e.g., Qualtrics) and web-browsers.

Limitations of the reanalysis should be noted. It is not always clear from the reporting articles which test statistic is most suitable to extract for purposes of reanalysis. One main reason for this difficulty was outlined earlier in the methods section of the study – inconsistent reporting practices. Despite a clear and detailed article published in *American Psychologist* by the APA in 2008 that discusses desirable reporting standards in psychology, and other initiatives in other fields to improve research reporting (e.g., the guidelines developed to improve the reporting of randomized-controlled trials in health-related research: Moher, Schulz, & Altman, 2001), many researchers in the social sciences have failed to adopt them (Mayo-Wilson, 2013). To be clear, poor standards of reporting are not the norm only in psychological science. To illustrate: Mackey (2012) in linguistics research states that insufficient reporting of details important for replication is problematic in many studies (p. 26); Button and colleagues (2013) in biomedical research, discuss the relationship between insufficient reporting of statistical details and false positives in results. We also recognize that it is difficult to manage a good balance between adequate reporting and the word limit in many (especially higher-impact) journals. Though, on the other hand, authors can upload supplementary documents to the various platforms available (the Open Science Framework, or Curate Science, for instance), or submit.

Another limitation regards our reanalysis of only t-tests. While reanalysis of more complex designs is possible using the Bayes factor package, we only demonstrate with the simpler design of the t-test. We intend to show, by this demonstration, a proof of concept of a methodical and evidence-driven approach to choosing targets for replication. The Bayesian reanalysis is a clear strength, from which replicating labs can draw, however we do not advocate only the use of a Bayesian reanalysis. We must consider factors that place the article and its content in context. We must consider its appropriateness as a study for replication (is a replication feasible for less well-equipped or specialist labs?), as well as the literature body it is part of. Is the study generally well supported, or does it tell a story conflicting with existing findings? Is it theoretically important, and does it hold relevance in its current historical, social and cultural context?

The reader may wonder why we have chosen not to assess the soundness of certain aspects of the methodologies of the original studies as a criterion for what studies to replicate. Although we argue that such a set of assessments is outside of the scope of the article, we recognize that to attempt to replicate an effect elicited by a poor methodological set-up is ill advised. We recommend that users of our method use their own judgment to determine whether or not an original article's methods are sound, and to consider each experiment of their final filtered sample in turn. If the methods of the final sample of potential targets is difficult for a user to assess (for example, perhaps one ends up with two targets using highly technical methods that the typical user may be unfamiliar with), the user may want to limit themselves to those studies for which they are confident assessing the soundness of the chosen methodology.

A practical yet somewhat philosophical argument must be raised of how one might use the Bayesian reanalysis to *prioritize* replication targets. The reader critical of Bayes factors may suggest that no matter what classification one uses (Jeffreys or otherwise), Bayes factors still do not provide a complete measure of the information contained in a given original study. This reader would be right, though this can be said for any currently used quantification approach. We stress that we are not advertising the Bayesian reanalysis as the *only* route to a search for replication targets. We argue that it is a tool one can apply to reveal valuable information to use to distinguish between pro-null evidence and ambiguous study results. In this demonstration, it was valuable as a kind of centrifuge – filtering the studies into different 'weight' categories based on the evidence from the results, which helps us determine which studies should be replicated first. The Bayesian reanalysis can be conducted relatively easily for most interested users with the statistical software R, using the code we have provided on our OSF page https://doi.org/10.17605/OSF.IO/3RF8B, to reduce the amount of potential replication targets, allowing individuals to direct their resources in a manner based on a justifiable and systematic method.

In this paper, we have chosen to have our statistical considerations be guided by the strength of evidence for the existence of an effect. Strong evidence can result from a large sample drawn from a relatively modest true effect, or a modest sample drawn from a large true effect. Other criteria are conceivable, such as those based on the precision of the effect size estimate.

A final important consideration for the reader concerns the role of publication bias in the pool of potential targets, and therefore final target selection. The work of Etz and Vandekerckhove (2016) suggests that if one were to take all studies as the possible pool of targets (that is, take publication bias into account), the average effect size will be smaller, and, presumably, the pool of viable targets much larger. Although their results suggest that an estimate of average strength of evidence based on published results is an overestimate, under the assumptions that (1) a single study has not been replicated many times in the same lab and only the most compelling result reported; and (2) a single study has not been duplicated exactly somewhere else in the world but was never reported; the reported test statistic can be safely reanalyzed in the way we have in our paper.

Aside from this, to date over 200 academic journals use the registered report format (for an up-to-date figure, see https://cos.io/rr/), and the number is steadily climbing. We consider it likely that as time passes and more people take advantage of this submission format, publication bias prevalence will decrease.

We would like to stress that the articles discussed in detail in this study were selected for illustration purposes only. The demonstration serves as proof of concept, and by no means aims to criticize specific studies or question their veracity. In fact, one of the three articles has two OSF badges (for more information see https://cos.io/our-services/open-science-badges-details/): one for open data, and one for open materials, indicating that the authors have made their data and study materials openly available on their project's OSF page. One of the other articles has the badge for open materials. The third article has provided access to their study materials in a supplemental folder available on the *Psychological Science* website. Such a commitment to transparent scientific practices are associated with research that is of higher quality, and therefore likely to be more reproducible (see the OSF badge page: https://cos.io/our-services/open-science-badges-details/ for a discussion).

The current debate over poor reproducibility in psychology has led to a number of new ideas for how to improve our research going forward. Increased numbers of replication studies is one such advance, which has been taken up wholeheartedly by many concerned researchers. While such an initiative marks a positive and constructive move toward remedying a serious problem in our field, it is neither efficient nor useful to replicate results randomly. In this article, we have argued for and demonstrated an approach which is methodical and systematic, supplemented by careful and defensible qualitative analysis toward the selection of replication targets.

The approach we advocate and apply in this article can be simple and relatively fast to conduct, and affords the user access to important information about the strength of evidence contained in a published study. Although efficient, this approach has the potential to maximize the impact of the outcomes of those replications, and minimize the waste of resources that could result from a haphazard approach to replication. Combining a quantitative reanalysis with a qualitative assessment process of a large group of potential replication targets in a simple approach such as the one presented in this paper, allows the information of multiple sources to prioritize replication targets, and can assist in refining the methodology of the replication study.

Appendix A

Table showing details of each reanalyzed result, and relevant information associated with each article. A full spreadsheet of all information can be found at the project's OSF page https://doi.org/10.17605/OSF.IO/3RF8B.

Article	Result	Authors	Year	T	DF	Overall N	p-value reported	BF(10)	Evidence Tier
1	a	Ding et al.	2015	4.42	40	42	<.001	283.12	1
1	b	Ding et al.	2015	3.49	40	42	<.001	27	1
2		Metcalfe et al.	2015	7.28	87	89	<.0001	106765637.21	1
4	a	Reinhart et al.	2015	2.605	17	18	0.018	3.19	1
5	a	Fan et al.	2015	2.81	46	48	0.007	6.25	1
5	b	Fan et al.	2015	2.51	46	48	0.016	3.44	1
6	a	Schroeder et al.	2015	3.79	157	160	<.01	213	1
7	a	Mackey et al.	2015	4.4	56	58	0.0001	412.32	1
7	b	Mackey et al.	2015	4.7	56	58	<.0001	1030.14	1
8	a	Dai et al.	2015	2.47	214	216	0.01	5.05	1
9	a	Okonofua et al.	2015	4.06	23	25	<.001	139.62	1
9	b	Okonofua et al.	2015	–4.99	23	25	<.001	1158.7	1
10	b	Olson et al.	2015	4.3	16	17	0.001	126.81	1
10	a	Olson et al.	2015	3.89	29	30	0.001	114.57	1
10	c	Olson et al.	2015	6.75	29	30	<.001	151537.61	1
11	a	Yin et al.	2015	5.73	15	16	<.001	644.57	1
11	b	Yin et al.	2015	3.23	15	16	0.006	8.84	1
11	d	Yin et al.	2015	5.88	15	16	<.001	1646.23	1
11	e	Yin et al.	2015	2.59	15	16	0.021	6	1
11	f	Yin et al.	2015	2.84	15	16	0.012	9.07	1
14		Storm et al.	2015	3.23	19	20	0.004	10.21	1
15		Perilloux et al.	2015	7.49	482	484	<.001	18931144326.12	1
18		Porter et al.	2016	2.89	85	88	0.005	7.91	1

(Contd.)

Article	Result	Authors	Year	T	DF	Overall N	p-value reported	BF(10)	Evidence Tier
19		Skinner et al.	2016	4.25	66	67	<.001	297.55	1
20		Kirk et al.	2016	3.59	43.35	54	0.001	40.35	1
22	a	Cooney et al.	2016	3.76	29	30	0.001	83.98	1
22	b	Cooney et al.	2016	4.27	57	59	<.001	285.37	1
22	c	Cooney et al.	2016	6.83	149	150	<.001	42432905.55	1
23		Zhou et al.	2016	7.26	70	73	<.001	19638415.24	1
25	b	Saint-Aubin et al.	2016	6.02	34	35	<.0001	21066.77	1
25	a	Saint-Aubin et al.	2016	5.6	45	46	<.0001	13805.45	1
26	a	Li et al.	2016	4.08	22	24	0.0005	53.71	1
26	b	Li et al.	2016	3.86	26	28	0.00068	38.61	1
29		Sloman et al.	2016	-3.4	69	70	0.001	22.86	1
30	b	Picci et al.	2016	2.8	27	28	0.001	10.5	1
30	c	Picci et al.	2016	4.4	27	28	0.001	273.96	1
30	d	Picci et al.	2016	3.14	29	30	0	20.56	1
3		Madore et al.	2015	2.49	22	23	0.021	2.67	3
4	b	Reinhart et al.	2015	2.318	17	18	0.033	1.99	3
4	c	Reinhart et al.	2015	2.326	17	18	0.033	2.02	3
4	d	Reinhart et al.	2015	2.263	17	18	0.04	1.83	3
4	e	Reinhart et al.	2015	2.466	17	18	0.027	2.53	3
6	b	Schroeder et al.	2015	2.09	215	218	0.04	1.14	3
8	b	Dai et al.	2015	2	211	213	0.047	0.97	3
11	c	Yin et al.	2015	2.47	15	16	0.026	2.52	3
12	a	Kupor et al.	2015	2.21	59	61	0.031	1.96	3
12	c	Kupor et al.	2015	2.22	98	100	0.029	1.83	3
12	b	Kupor et al.	2015	2.27	200	202	0.024	1.68	3
13		Farooqui et al.	2015	2.2	20	21	0.04	1.64	3
16		Olsson et al.	2016	2.44	97	100	0.02	2.85	3
17		Watson-Jones et al.	2016	2.05	86	88	0.043	1.38	3
21		Hung et al.	2016	-2.51	19	20	0.02	2.75	3
24	b	Hsee et al.	2016	2.35	17	20	<.031	2.37	3
24	a	Hsee et al.	2016	2.25	52	54	0.029	2.13	3
27		Constable et al.	2016	2.1	35	38	0.04	1.7	3
28		Chen et al.	2016	2.39	187	189	0.018	2.21	3
30	a	Picci et al.	2016	2.25	29	30	0.032	2.15	3

Data Accessibility Statement

The database including all article information and reanalyzed Bayes factors are available, along with the analysis and plot R scripts, on the project's OSF page: https://doi.org/10.17605/OSF.IO/3RF8B.

Supplemental Material

All files associated with this study are found on the project's OSF page: https://doi.org/10.17605/OSF.IO/3RF8B

Notes

[1] For a more detailed primer on the Bayes factor, please see Appendix A in Field and colleagues (2016); for a full expose, see Etz and Vandekerckhove (2018).

[2] We note that some of Mackey's guidelines lead to subjective decisions about what is theoretically relevant and important. What may be theoretically important in one field, may not be worth investigating in another, and so it is vital to consider the context of a potential

replication target, and root one's judgments in quantifiable argumentation.

[3] Bayes factors can show evidential strength in favor of an alternative hypothesis (denoted BF_{10}), or be inverted and show support for the null hypothesis (denoted BF_{01}). In this article, we only discuss Bayes factors in terms of their support of the alternative, and so refrain from using the specific subscript notation or verbal indication.

[4] We originally planned to consider those articles with the smallest Bayes factors, however, as we discuss later, there are many results with similar Bayes factors (e.g., 1.64, 1.68 and 1.70), which makes that choice alone somewhat arbitrary.

[5] More complicated approaches to handle the case of multiple studies in a single paper corroborating a certain claim in the manuscript exist, for instance through a Bayesian model-averaged meta-analysis.

[6] We only target these articles to practically demonstrate how our approach can be used. We do not imply that they are of low veracity or that the results were obtained by questionable means.

[7] Of course, the replication as described here would need to feature different risk-taking activities, as aged persons may be averse in general to activities such as skydiving.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

- DvR and RH conceived of the idea of reanalyzing Bayes factors to quantify evidential strength of original article results; SMF conceived of the qualitative analysis, and overall process
- SMF extracted all article information which formed the data file analyzed in the study
- DvR wrote the code for the reanalysis phase. SMF analyzed and interpreted the findings derived from it; DvR, RH and LB refined the interpretations and plots for the final manuscript
- SMF drafted the article; SMF, DvR, RH and LB further revised it
- SMF approved the submitted version for publication

References

Aczel, B., Palfi, B., & **Szaszi, B.** (2017). Estimating the evidential value of significant results in psychological science. *PLOS ONE*, *12*, e0182651. DOI: https://doi.org/10.1371/journal.pone.0182651

Aschwanden, C. (2016, March 24). *Failure is moving science forward*. FivethirtyEight. Retrieved from https://fivethirtyeight.com/features/failure-is-moving-science-forward/

Baker, M. (2015, August 27). *Over half of psychology studies fail reproducibility test*. Nature News. Retrieved from https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248. DOI: https://doi.org/10.1038/nature.2015.18248

Baker, M. (2016, February 4). *Biotech giant publishes failures to confirm high-profile science*. Nature News. Retrieved from https://www.nature.com/news/biotech-giant-publishes-failures-to-confirm-high-profile-science-1.19269. DOI: https://doi.org/10.1038/nature.2016.19269

Begley, C. G., & **Ioannidis, J. P.** (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*, 116–126. DOI: https://doi.org/10.1161/CIRCRESAHA.114.303819

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & **Munafò, M. R.** (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. DOI: https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., ..., Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. DOI: https://doi.org/10.1038/s41562-018-0399-z

Chang, A. C., & **Li, P.** (2018). Is economics research replicable? Sixty published papers from thirteen journals say "often not". *Critical Finance Review*, *7*, 1–25.

Chawla, D. S. (2016, February 26). *How many replication studies are enough?* Nature News. Retrieved from https://www.nature.com/news/how-many-replication-studies-are-enough-1.19461

Clarke, A. D., Barr, C., & **Hunt, A. R.** (2016). The effect of visualization on visual search performance. *Attention, Perception, & Psychophysics*, *78*, 2357–2362. DOI: https://doi.org/10.3758/s13414-016-1174-8

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & **Lakens, D.** (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, e124. DOI: https://doi.org/10.1017/S0140525X18000596

Dablander, F. (2017, May 25). *Are you registering that? an interview with prof. chris chambers*. JEPS Bulletin. Retrieved from https://blog.efpsa.org/2017/05/25/are-you-registering-that-an-interview-with-prof-chris-chambers/

Dachtler, J., & **Fox, K.** (2017). Do cortical plasticity mechanisms differ between males and females? *Journal of Neuroscience Research*, *95*, 518–526. DOI: https://doi.org/10.1002/jnr.23850

Dai, H., Mao, D., Riis, J., Volpp, K. G., Relish, M. J., Lawnicki, V. F., & **Milkman, K. L.** (2017). Effectiveness of medication adherence reminders tied to "fresh start" dates: A randomized clinical trial. *Jama Cardiology*, *2*, 453–455. DOI: https://doi.org/10.1001/jamacardio.2016.5794

Dai, H., Milkman, K. L., & **Riis, J.** (2014). The fresh start effect: Temporal landmarks motivate aspirational behavior. *Management Science*, *60*, 2563–2582. DOI: https://doi.org/10.1287/mnsc.2014.1901

Dai, H., Milkman, K. L., & **Riis, J.** (2015). Put your imperfections behind you: Temporal landmarks spur goal initiation when they signal new beginnings.

Psychological Science, *26*, 1927–1936. DOI: https://doi.org/10.1177/0956797615605818

Etz, A., & **Vandekerckhove, J.** (2016). A Bayesian perspective on the reproducibility project: Psychology. *PloS one*, *11*, e0149794. DOI: https://doi.org/10.1371/journal.pone.0149794

Etz, A., & **Vandekerckhove, J.** (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. DOI: https://doi.org/10.3758/s13423-017-1262-3

Field, S. M., Wagenmakers, E.-J., Newell, B. R., Zeelenberg, R., & **van Ravenzwaaij, D.** (2016). Two Bayesian tests of the GLOMOsys Model. *Journal of Experimental Psychology: General*, *145*, e81. DOI: https://doi.org/10.1037/xge0000067

Green, C. S., & **Bavelier, D.** (2012). Learning, attentional control, and action video games. *Current Biology*, *22*, 197–206. DOI: https://doi.org/10.1016/j.cub.2012.02.012

Grimes, D. R., Bauch, C. T., & **Ioannidis, J. P.** (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*, Article 171511. DOI: https://doi.org/10.1098/rsos.171511

Gronau, Q. F., Ly, A., & **Wagenmakers, E.-J.** (2019). Informed Bayesian t-tests. *The American Statistician*, 1–14. DOI: https://doi.org/10.1080/00031305.2018.1562983

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & **Wagenmakers, E. J.** (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138. DOI: https://doi.org/10.1080/23743603.2017.1326760

Hackett, C., Stonawski, M., Potancoková, M., Grim, B. J., & **Skirbekk, V.** (2015). The future size of religiously affiliated and unaffiliated populations. *Demographic Research*, *32*, 829–842. DOI: https://doi.org/10.4054/DemRes.2015.32.27

Hardwicke, T. E., Tessler, M. H., Peloquin, B. N., & **Frank, M. C.** (2018). A Bayesian decision-making framework for replication. *Behavioral and Brain Sciences*, *41*, e132. DOI: https://doi.org/10.1017/S0140525X18000675

Harris, C. (2018, March 21). Young people in UK and Netherlands among Europe's least religious. *EuroNews*. Retrieved from http://www.euronews.com/2018/03/21/how-europe-s-young-adults-are-turning-their-backs-on-religion

Haucke, M., Miosga, J., Hoekstra, R., & **van Ravenzwaaij, D.** (2019). Bayesian frequentists: Examining the paradox between what researchers can conclude versus what they want to conclude from statistical results. Manuscript submitted for publication.

Hoekstra, R., Monden, R., van Ravenzwaaij, D., & **Wagenmakers, E.-J.** (2018). Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. *PloS one*, *13*, e0195474. DOI: https://doi.org/10.1371/journal.pone.0195474

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*, e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

Jeffreys, H. (1955). The present position in probability theory. *The British Journal for the Philosophy of Science*, *5*, 275–289. DOI: https://doi.org/10.1093/bjps/V.20.275

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

John, L. K., Loewenstein, G., & **Prelec, D.** (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. DOI: https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., ..., Nosek, B. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. DOI: https://doi.org/10.1177/2515245918810225

Kuehberger, A., & **Schulte-Mecklenbeck, M.** (2018). Selecting target papers for replication. *Behavioral and Brain Sciences*, *41*, e139. DOI: https://doi.org/10.1017/S0140525X18000742

Kupor, D. M., Laurin, K., & **Levav, J.** (2015). Anticipating divine protection? Reminders of God can increase nonmoral risk taking. *Psychological Science*, *26*, 374–384. DOI: https://doi.org/10.1177/0956797614563108

Lee, J., & **Dai, H.** (2017). The motivating effects of temporal landmarks: Evidence from the field and lab. *Missouri Law Review*, *82*, 683–694. DOI: https://doi.org/10.1201/b12494-35

Mackey, A. (2012). Why (or why not), when and how to replicate research. In: G. Porte (Ed.), *Replication research in applied linguistics* (pp. 34–69). Cambridge, UK: Cambridge University Press.

Makel, M. C., Plucker, J. A., & **Hegarty, B.** (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542. DOI: https://doi.org/10.1177/1745691612460688

Marszalek, J. M., Barber, C., Kohlhart, J., & **Cooper, B. H.** (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*, 331–348. DOI: https://doi.org/10.2466/03.11.PMS.112.2.331-348

Mogilner, C., Hershfield, H. E., & **Aaker, J.** (2018). Rethinking time: Implications for well-being. *Consumer Psychology Review*, *1*, 41–53. DOI: https://doi.org/10.1002/arcp.1003

Moher, D., Schulz, K. F., & **Altman, D. G.** (2001). The consort statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology*,

1, 657–662. DOI: https://doi.org/10.1186/1471-2288-1-2

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., ..., Zwaan, R. A. (2016). The peer reviewers' openness initiative: incentivizing open research practices through peer review. *Royal Society Open Science*, *3*, Article 150547. DOI: https://doi.org/10.1098/rsos.150547

Morey, R. D., Rouder, J. N., & **Jamil, T.** (2015). Package 'BayesFactor' [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., ..., Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. DOI: https://doi.org/10.1126/science.aab2374

OSC. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 947–951.

Peetz, J., & **Wilson, A. E.** (2013). The post-birthday world: Consequences of temporal landmarks for temporal self-appraisal and motivation. *Journal of Personality and Social Psychology*, *104*, 249–267. DOI: https://doi.org/10.1037/a0030477

Pittelkow, M., Hoekstra, R., & **van Ravenzwaaij, D.** (2019). Preliminary evidence for the replication crisis in clinical psychology: A Bayesian and theoretical re-evaluation of the evidence in published literature. *Manuscript submitted for publication.*

Reinhart, R. M., McClenahan, L. J., & **Woodman, G. F.** (2015). Visualizing trumps vision in training attention. *Psychological Science*, *26*, 1114–1122. DOI: https://doi.org/10.1177/0956797615577619

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. DOI: https://doi.org/10.1037/0033-2909.86.3.638

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & **Iverson, G.** (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

Rutledge, R. B., Smittenaar, P., Zeidman, P., Brown, H. R., Adams, R. A., Lindenberger, U., ..., Dolan, R. J. (2016). Risk taking for potential reward decreases across the lifespan. *Current Biology*, *26*, 1634–1639. DOI: https://doi.org/10.1016/j.cub.2016.05.017

Schimmack, U. (2018). The replicability revolution. *Behavioral and Brain Sciences*, *41*, e147. DOI: https://doi.org/10.1017/S0140525X18000833

Schneiders, B. (2013, November 5). Growing numbers of young Australians record no religion in census. *The Sydney Morning Herald*. Retrieved from https://www.smh.com.au/national/growing-numbers-of-young-australians-record-no-religion-in-census-20131124-2y3w7.html

Urminsky, O. (2017). The role of psychological connectedness to the future self in decisions over time. *Current Directions in Psychological Science*, *26*, 34–39. DOI: https://doi.org/10.1177/0963721416668810

Verhagen, A. J., & **Wagenmakers, E.-J.** (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology, General*, *143*, 1457–1475. DOI: https://doi.org/10.1037/a0036731

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p*-values. *Psychonomic Bulletin & Review*, *14*, 779–804. DOI: https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & **Forstman, B.** (2014). Rewarding high-power replication research. *Cortex*, *51*, 105–106. DOI: https://doi.org/10.1016/j.cortex.2013.09.010

Wang, A. (2015, November 4). Today's young Americans are less religious – and a lot more likely to stay that way. *Quartz*. Retrieved from https://qz.com/540395/todays-young-americans-are-less-religious-and-a-lot-more-likely-to-stay-that-way/

Wetzels, R., Matzke, D., Michael, M. D., Rouder, J. N., Iverson, G. J., & **Wagenmakers, E.-J.** (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t*-tests. *Perspectives on Psychological Science*, *6*, 291–298. DOI: https://doi.org/10.1177/1745691611406923

Witte, E. H., & **Zenker, F.** (2018). Data replication matters to an underpowered study, but replicated hypothesis corroboration counts. *Behavioral and Brain Sciences*, *41*, e156. DOI: https://doi.org/10.1017/S0140525X18000924

Woolston, C. (2015). Online debate erupts to ask: Is science broken? *Nature*, *519*, 393. DOI: https://doi.org/10.1038/519393f

Wu, E. C., & **Cutright, K. M.** (2018). In God's hands: How reminders of God dampen the effectiveness of fear appeals. *Journal of Marketing Research*, *55*, 119–131. DOI: https://doi.org/10.1509/jmr.15.0246

Yotsumoto, Y., Chang, L.-H., Ni, R., Pierce, R., Andersen, G. J., Watanabe, T., & **Sasaki, Y.** (2014). White matter in the older brain is more plastic than in the younger brain. *Nature Communications*, *5*, 5504. DOI: https://doi.org/10.1038/ncomms6504

Zellner, A., & **Siow, A.** (1980). Posterior odds ratios for selected regression hypotheses. In: J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press. DOI: https://doi.org/10.1007/BF02888369

Zwaan, R. A., Etz, A., Lucas, R. E., & **Donnellan, M. B.** (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*. DOI: https://doi.org/10.1017/S0140525X17001972

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.218.pr

How to cite this article: Field, S. M., Hoekstra, R., Bringmann, L., & van Ravenzwaaij, D. (2019). When and Why to Replicate: As Easy as 1, 2, 3? *Collabra: Psychology,* 5(1): 46. DOI: https://doi.org/10.1525/collabra.218

Senior Editor: Victoria Savalei

Editor: Victoria Savalei

Submitted: 18 December 2018 **Accepted:** 14 September 2019 **Published:** 30 September 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS